Information Interaction

Thoughts on the intersection of user experience, search, text analytics & more

Feeds: Posts Comments

Zipf’s Law: how can something so simple explain so much

November 8, 2010 by Tony Russell-Rose

One of the threads emerging from the recent Search at the Guardian (https://isquared.wordpress.com/2010/10/20/search-at-the-guardian-newspaper/) event was how difficult it is to get search right, particularly when so many of the queries in a typical day belong to a long tail (http://en.wikipedia.org/wiki/Long_Tail) of relatively unique “edge cases”. In effect, there are a small number queries that are incredibly common, but a vast number of queries that are incredibly rare. Optimising for the head is an ostensibly straightforward undertaking; optimising for the tail an entirely different proposition.

But this phenomenon isn’t unique to the Guardian – in fact it is true of language in general: almost all* natural languages are characterised by Zipf’s Law (http://en.wikipedia.org/wiki/Zipf%27s_Law), which states that “*given some corpus (http://en.wikipedia.org/wiki/Text_corpus) of natural language (http://en.wikipedia.org/wiki/Natural_language) utterances, the frequency of any word is inversely proportional (http://en.wikipedia.org/wiki/Inversely_proportional) to its rank*“. So the most common word will occur twice as often as the second ranked word, etc. And it’s not just the terms themselves – if you take higher-order structures, such as named entities (http://en.wikipedia.org/wiki/Named_entity_recognition#Named_entity_types), part of speech tags (http://en.wikipedia.org/wiki/Part-of-speech_tagging), partial parses (http://en.wikipedia.org/wiki/Shallow_parsing), etc. you find the same phenomenon: a power law (http://en.wikipedia.org/wiki/Power_law) probability distribution. (http://en.wikipedia.org/wiki/Probability_distribution) Intriguingly, the same relationship occurs in many other phenomena totally unrelated to language, such as the population ranks of cities in various countries, corporation sizes, income rankings, and so on.

But I digress. The reason I mention this is that the same issue we saw at the Guardian a couple of weeks ago was fundamental to our work at Reuters (http://www.reuters.com/) almost a decade ago. We may have had a different set of goals back then, but the underlying principles were the same. In our case, we used the insight into those distributions to inform our approach to the creation and management of taxonomies (and associated linguistic resources).

By good fortune, we documented our initial work on this in the short article below. Almost a decade on, it's still worth reading verbatim. BTW this work was done in collaboration with my ex-Reuters colleague Mark Stevenson (http://staffwww.dcs.shef.ac.uk/people/M.Stevenson/).

*NB Ever since my PhD days I'd always believed that Zipf's Law was a universal, i.e. something innate to the structure of all human languages. But according to wikipedia, it applies only to "most". Can anyone cite a counter-example?

Finding Structure in Unstructured Data

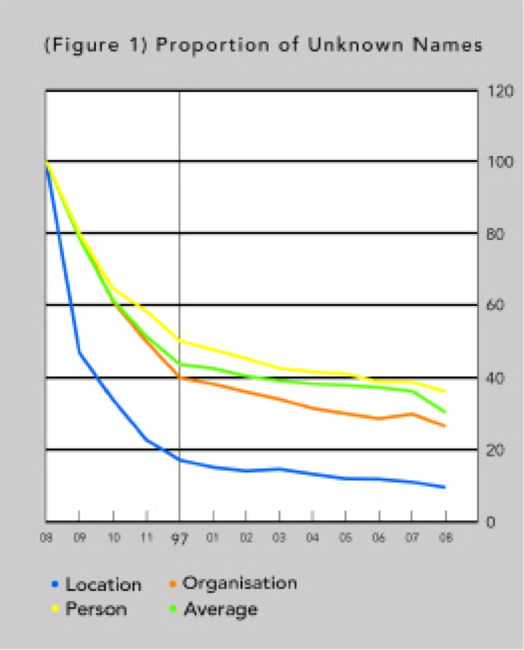
Reuters produces 11,000 news stories daily, for a variety of financial and media clients. Many people rely on this information for making key business decisions. An important part of that decision-making process is 'contextualizing' the news – i.e. relating it to other information, such as stock quotes, investment research, company reports, and so on. So how can companies like Reuters help their clients gain access to such related information?

One way is through the construction and maintenance of information directories, such as catalogues of people, places, and organisations. A database of people, for example, could provide biographical details of all the key individuals in the news. Likewise, a database of organisations could provide valuable historical and reference data. And a database of places could give geographic locations, map references, population data, and so on.

But building and maintaining such directories is clearly a significant task for all but the most limited of applications. Besides, how large does such a directory need to be? How many key individuals does a company like Reuters need to include to provide 90% coverage (for example) of all the people mentioned in an average day's news? Similarly, how many companies are needed to give 90% coverage of all the organisations mentioned?

To answer these questions, Reuters Chief Technology Office (CTO) turned to InXight (http://en.wikipedia.org/wiki/Inxight)'s ThingFinder (http://inxightfedsys.com/products/sdks/tf/), and a freely-available collection of news stories known as the "Reuters Corpus (http://about.reuters.com/researchandstandards/corpus/)". This collection contains 800,000+ Reuters news stories, from August 1996 to August 1997 – a full calendar year of (English language) Editorial output.

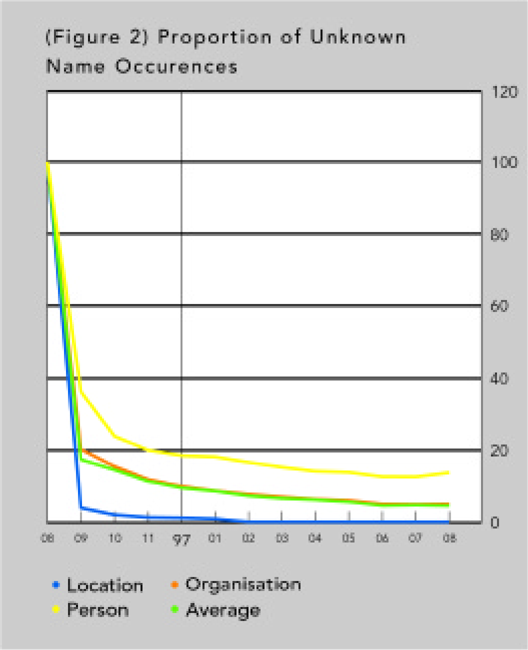
CTO staff set to work on the problem by trying to measure the rate at which names (of people, places and organisations) appeared over the 12 months of the Reuters Corpus. Each month's data was analysed in turn and the percentage of unique names that had not been seen in any previous month was calculated. Of course, during the first month, all the names were new, so the proportion of unknown names was 100%. But over the course of the year's data, month by month, more names could be added to the list of known (i.e. previously seen) names, so the proportion of unknown (i.e. new) names became progressively lower. However, even after gathering 11 months of data, it became clear that the proportion of unknown names would take much, much longer to approach zero (if indeed it ever would). In fact, by the 12th month, as many as 29.1% of the names had not occurred in the previous 11 months (see Figure 1).



(Figure 1) Proportion of Unknown Names

(https://isquared.files.wordpress.com/2010/11/fig1.png)Moreover, some types of name showed a higher 'turnover' (i.e. rate of change) than others: by the 12th month, only 7.3% (189) of locations were unknown, whereas 25.9% (6,038) of organisations were unknown, and as many as 33.4% (9,892) of people were unknown. In effect, this means that even if you gather data on all the people mentioned in the previous 11 months of data, this will still only cover around two-thirds of the people mentioned in your next month's output. Clearly, this has significant consequences for the prospects of building a comprehensive people directory.

However, there is one subtle distinction we should consider. The above analysis was based on the proportion of unique names, and ignores the number of times each name is mentioned (for example, Microsoft and Bill Clinton are two unique names, but both may be mentioned many hundreds of times in an average month). So if, by contrast, we perform the same analysis based on the actual occurrences of each name (and take into account such multiple counts), a different picture emerges – and we find now that only 5.6% of the names mentioned in the 12th month are new (see Figure 2).



(https://isquared.files.wordpress.com/2010/11/fig2.png)Moreover, when we break this figure down into the distribution for each name type, we find now that only 0.2% of locations are new, 5.4% of organisations are new, and 13.0% of people are new names. Evidently, in any one month, there are a few names that are mentioned a great many times, and a great many names that are mentioned only a few times. Interestingly, this effect is not restricted to named entities – indeed, the everyday words of most human languages also display this characteristic (a phenomenon known as 'Zipf's Law (http://en.wikipedia.org/wiki/Zipf%27s_law)').

These figures provide valuable evidence regarding the magnitude of effort involved in creating a comprehensive database of names (particularly of people). Clearly, providing complete coverage is almost impossible, since new names appear so frequently (as many as 29% in a given month will be unseen in the previous 11 months). On the other hand, since many of the actual name occurrences are accounted for by a smaller number of individuals, it may be possible to build a database that covers a large proportion of people occurrences based on just 11 months of data. Of course, the issue of exactly what level of coverage to provide remains to be decided. But based on the above analysis we can now plan for such an undertaking much more effectively.

Medical
Negligence
Patient Claim Line

Are You Eligible For A Claim? Fi
Today By Taking Our 30 Second
Test

OPEN

REPORT THIS AD

Posted in Information architecture, Metadata, Search, Text analytics | Tagged Information Retrieval, natural language processing, Reuters, Text analytics | 2 Comments

2 Responses

MisterP *on November 16, 2010 at 10:29 pm* | *Reply*

Is the Zipf Law more amazing than Gaussian distribution to describe natural and/or large human-related phenomenon?…

Tony Russell-Rose *on November 16, 2010 at 10:49 pm* | *Reply*

That's an interesting way of looking at it. One one level, maybe it isn't… but I think the key difference is that the Gaussian is so ubiquitous we almost expect to see it extend to language-related phenomena (or at least, my students invariably did when I taught NLP classes many years ago). So there must be something hard-wired into us that makes the power law distribution a more effective way of constructing natural languages (of almost any origin). I'd love to know why that is.

Comments RSS